Ewaluacja embeddingów modelu CLIP

Marcin

20 lutego 2025

Wprowadzenie

- Celem prezentacji jest ocena fine-tuningu modelu **CLIP**. Fine-tuning przeprowadzono na danych którymi były pary typu (*tekst*, *obraz*), gdzie *tekst* opisuje *obraz*.
- Porównanie par typu (*tekst*, *obraz*) jako embeddingów przed i po fine-tuningu.
- Wykorzystano metody do oceny jakości modelu:
  - Macierz podobieństw kategorii *(Category Similarity Matrix)*
  - Macierz korelacji centroidów *(Centroid Correlation Matrix)*
  - PodobieństwoTanimoto *(Tanimoto Similarity)*

Model i fine-tuning

Model:

- Użyto modelu CLIP: `CLIPModel.from_pretrained("openai/clip-vit-base-patch32")`
- Fine-tuning z konfiguracją LoRA
- Strata: Triplet Loss

Triplet Loss:

- Text-Image Loss: $\text{ReLU}(\text{sim}(t, n) - \text{sim}(t, p) + \text{margin})$
- Image-Image Loss: $\text{ReLU}(\text{sim}(a, n) - \text{sim}(a, p) + \text{margin})$
- $\text{sim}(x, y)$ to cosinusowa miara podobieństwa
- Całkowita strata: $\text{loss} = \text{loss\_text} + \alpha \cdot \text{loss\_img}$

Model i fine-tuning

Accuracy:

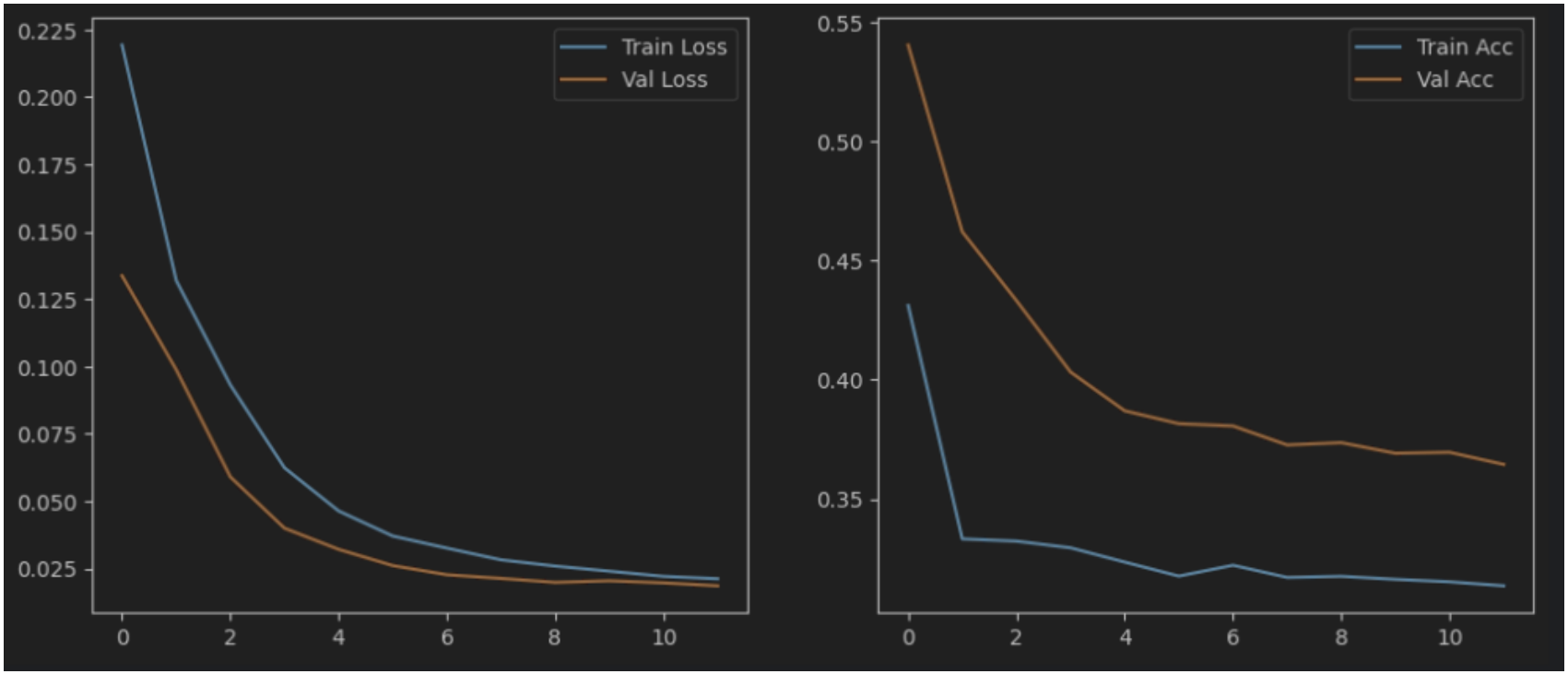
- Accuracy określa procent poprawnych dopasowań między tekstem a obrazem. Obliczana jest macierz podobieństwa kosinusowego między tekstami a obrazami. Predykcja to wybór obrazu o najwyższym podobieństwie dla danego tekstu. Accuracy to średnia poprawnych dopasowań

Parametry dla fine-tuningu:

- `BATCH_SIZE = 16`, `MARGIN = 0.3`, `EPOCHS = 12`

Dane: Fashion Product Text Images Dataset *(Kaggle)*
`https://www.kaggle.com/datasets/nirmalsankalana/fashion-product-text-images-dataset`
Wytrenowano na GPU: RTX 4000 Ada

Wykresy: Train/Validation Loss i Accuracy



Wykres train-loss, train-accuracy

Wprowadzenie - Cosine Similarity

Cosine similarity mierzy podobieństwo między dwoma wektorami A i B:

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

- Wartość bliska 1: wysokie podobieństwo.
- Wartość bliska -1: duże różnice między embeddingami.

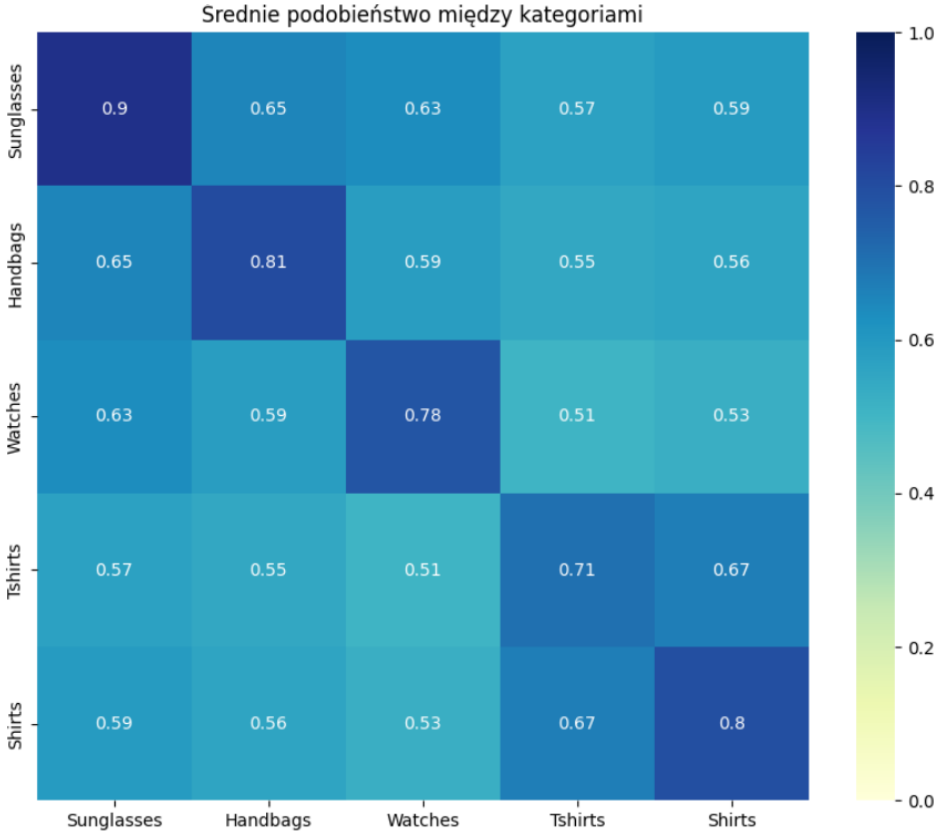
Obliczanie średnich

Średnia podobieństw dla par z tej samej kategorii (*avg_same*) oraz różnych kategorii (*avg_diff*):

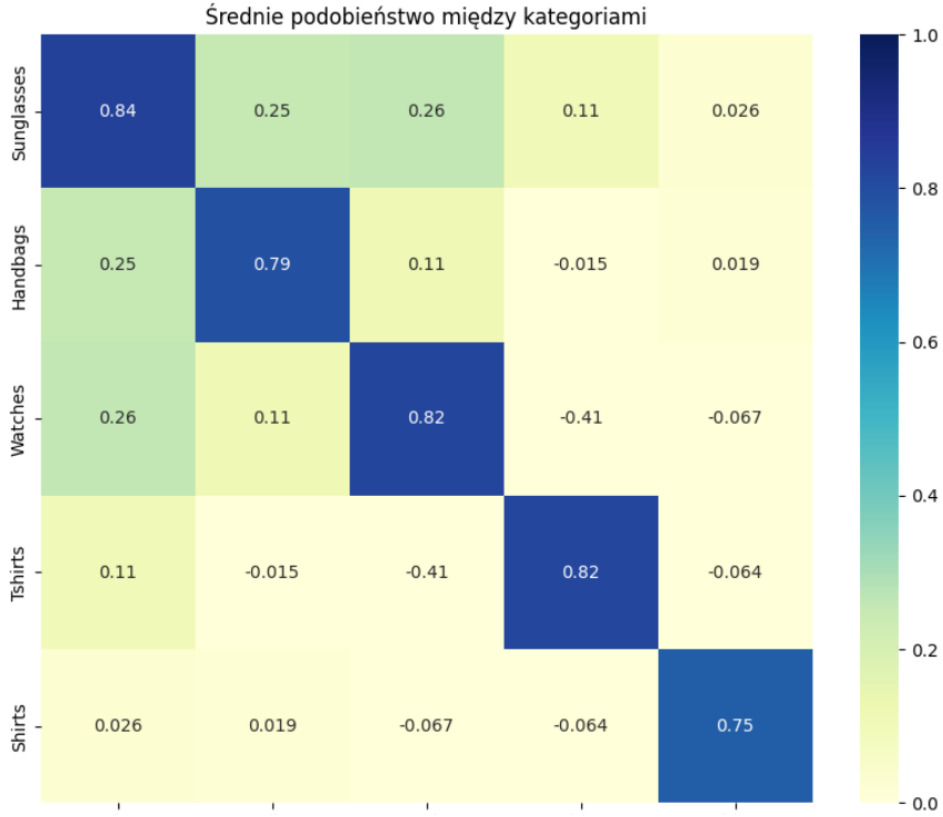
$$avg = \frac{1}{N} \sum_{i=1}^{N} \text{cosine\_similarity}(A_i, B_i) \tag{2}$$

Gdzie N to liczba par, a A_i i B_i to embeddingi par tekst-obraz.

Macierze podobieństw kategorii



Przed fine-tuning



Po fine-tuning

Macierz podobieństw kategorii po fine-tuningu - wnioski

- Fine-tuning znacząco poprawił zdolność modelu do rozróżniania kategorii.
- Macierz pokazuje większe podobieństwa wewnątrz kategorii i mniejsze między różnymi kategoriami.

Opis metody Centroid Correlation

- Celem metody jest obliczenie korelacji między centroidami kategorii na podstawie embeddingów CLIP.
- Używa metryki **cosine similarity** do oceny podobieństwa między centroidami kategorii.
- Metoda zwraca macierz korelacji oraz listę kategorii w odpowiedniej kolejności.

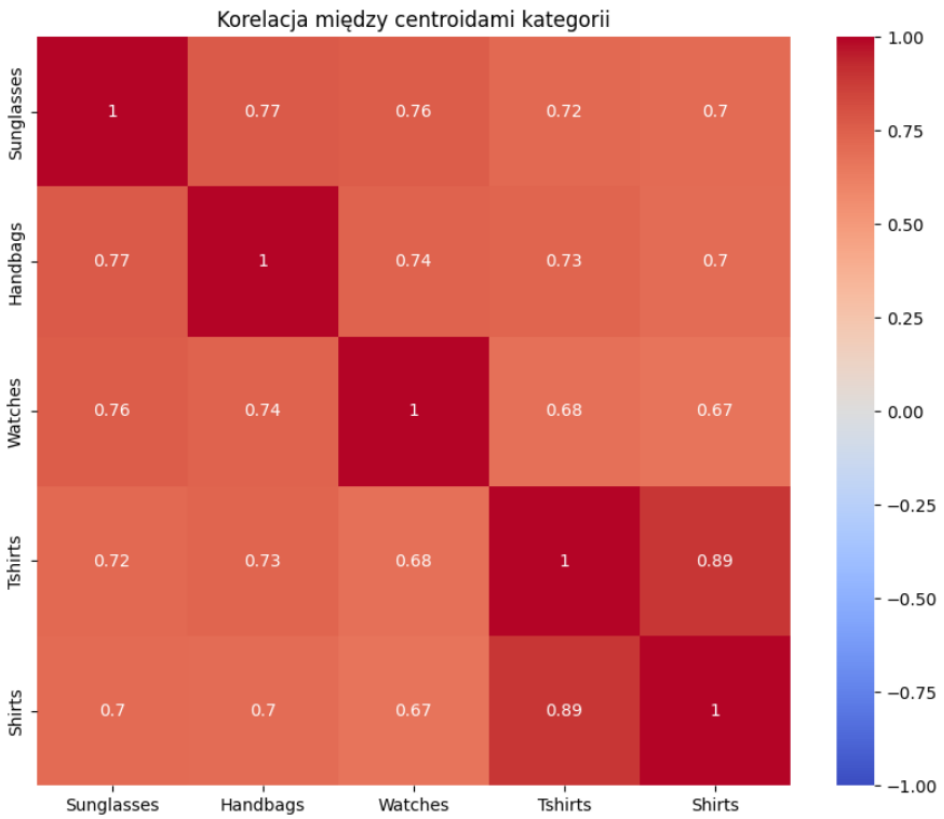
Cosine Similarity dla centroidów

Cosine similarity mierzy podobieństwo między dwoma wektorami centroidów C_1 i C_2:

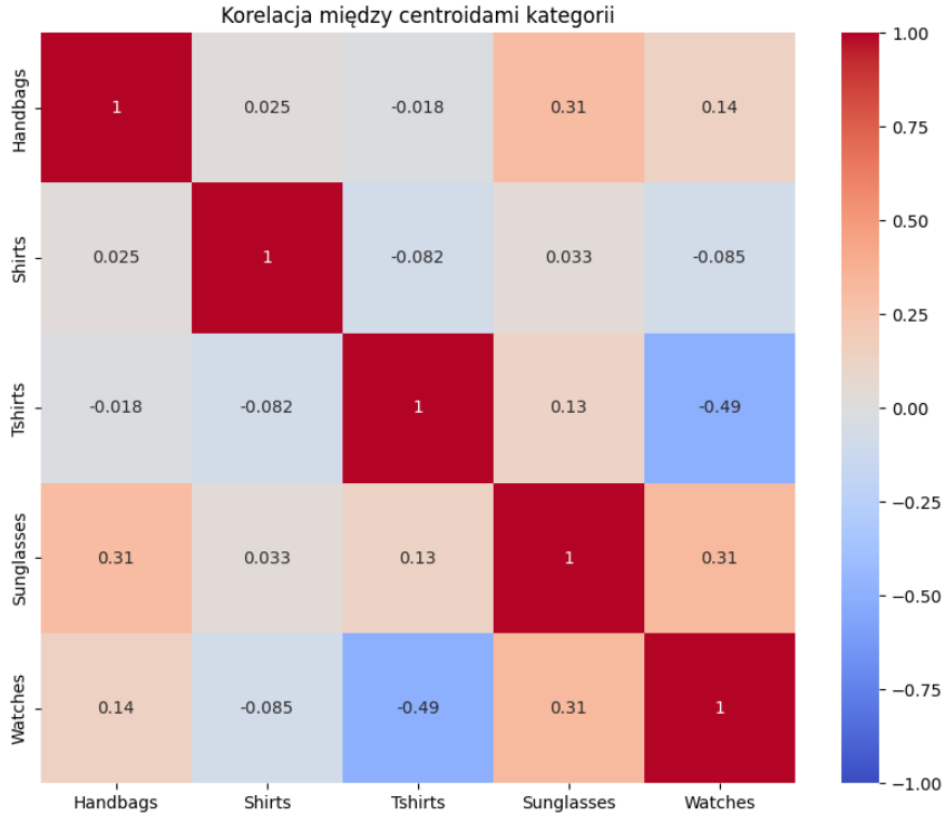
$$\text{cosine\_similarity}(C_1, C_2) = \frac{C_1 \cdot C_2}{\|C_1\| \|C_2\|} \tag{3}$$

- Centroid to średnia embeddingów w danej kategorii.
- Wartość 1 oznacza wysoką zgodność między kategoriami.
- Wartość -1 oznacza maksymalną różnicę między centroidami.

Macierz korelacji centroidów



Przed fine-tuning



Po fine-tuning

Macierz korelacji centroidów po fine-tuningu - wnioski

- Fine-tuning poprawił korelację między centroidami wewnątrz kategorii.
- Macierz pokazuje większe podobieństwa wewnątrz kategorii i mniejsze między różnymi kategoriami.

Opis metody wyznaczenia macierzy podobieństwa Tanimoto

- Metoda oblicza średnie podobieństwo Tanimoto między embeddingami dla różnych kategorii.
- Używa metryki **Tanimoto** do porównywania wszystkich możliwych par embeddingów między kategoriami.
- Zwraca macierz podobieństw oraz listę analizowanych kategorii.

Wzór na Tanimoto Similarity

Tanimoto similarity mierzy podobieństwo między wektorami a i b:

$$\text{tanimoto}(a, b) = \frac{a \cdot b}{\|a\|^2 + \|b\|^2 - a \cdot b} \tag{4}$$

- Wartość bliska 1 oznacza wysokie podobieństwo.
- Wartość 0 oznacza brak podobieństwa.

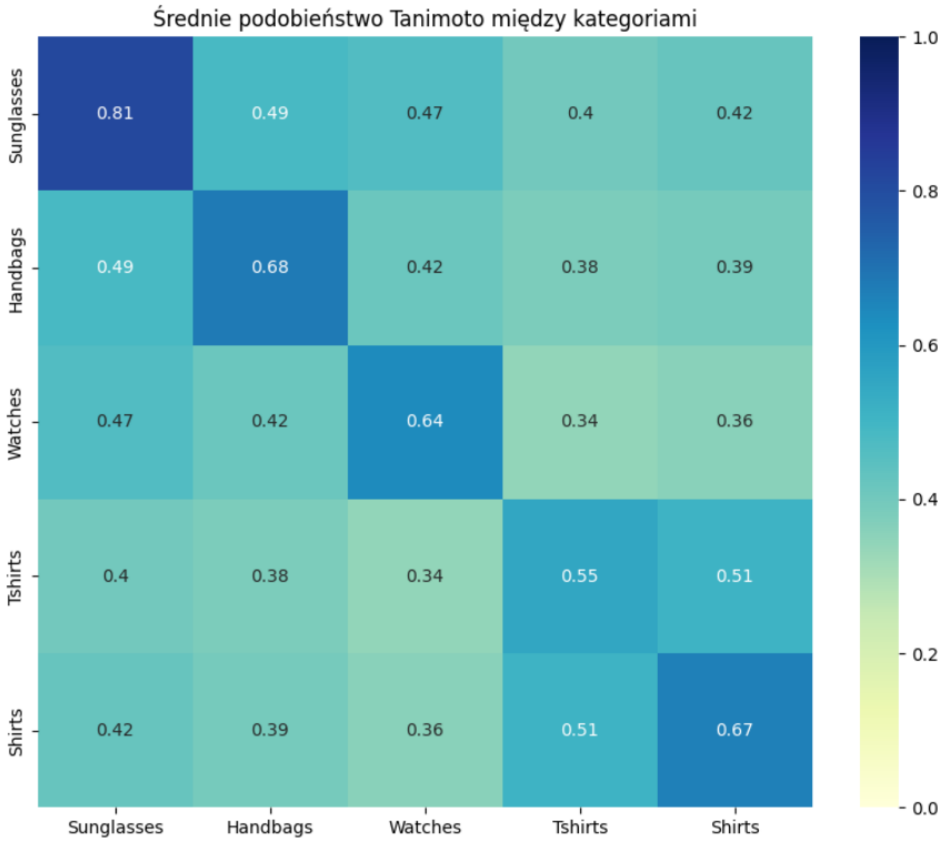
Tworzenie macierzy podobieństwa Tanimoto

Macierz podobieństwa Tanimoto tworzona jest przez obliczanie podobieństwa między każdą parą embeddingów w ramach wybranych kategorii. Dla każdej pary kategorii obliczane są wszystkie możliwe kombinacje embeddingów w danej kategorii oraz między kategoriami.

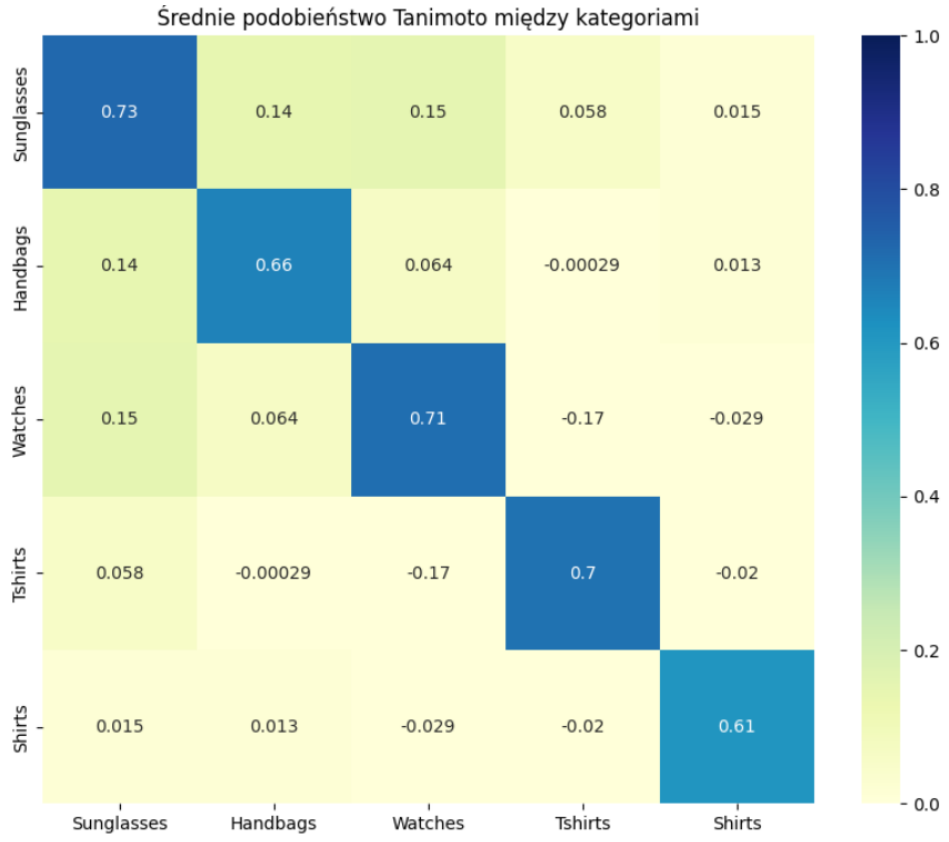
- Gdy kategorie są takie same, porównywane są tylko embeddingi wewnątrz tej samej kategorii.
- Gdy kategorie są różne, porównywane są wszystkie pary embeddingów między kategoriami.

W wyniku otrzymujemy macierz, w której elementy reprezentują średnie podobieństwo między kategoriami.

Macierz podobieństwa Tanimoto



Przed fine-tuning



Po fine-tuning

Macierz podobieństwa Tanimoto - po fine-tuningu - wnioski

- Macierz podobieństwa obrazuje przeciętne podobieństwo między embeddingami różnych kategorii.
- Przed fine-tuningiem podobieństwa mogą być niskie, szczególnie między różnymi kategoriami.
- Fine-tuning poprawia podobieństwa wewnątrz kategorii.
- Macierz ukazuje większą spójność embeddingów w ramach tej samej kategorii.

Dziękuję za uwagę!